DOCUMENT RESUME

ED 459 213 TM 033 512

AUTHOR Deng, Hui; Chang, Hua-Hua
TITLE A-Stratified Computerized Adaptive Testing with Unequal Item Exposure across Strata.
PUB DATE 2001-04-00
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Item Banks; *Selection; Simulation; *Test Items
IDENTIFIERS *Item Exposure (Tests); *Stratification

ABSTRACT

The purpose of this study was to compare a proposed revised a-stratified, or alpha-stratified, USTR method of test item selection with the original alpha-stratified multistage computerized adaptive testing approach (STR) and the use of maximum Fisher information (FSH) with respect to test efficiency and item pool usage using simulated computerized adaptive tests (CATs). Two simulation studies were conducted to compare the proposed USTR with its original STR version and FSH with respect to efficiency of trait estimation and item usage. In the first study, a set of item parameters were generated to form a bank of 400 items and the gains of using the USTR design were explored. In the second study, the comparison was carried out under a more complex design using a real item pool of 300 mathematical items. Results suggest that USTR holds considerable potential for improving test efficiency for STR. It effectively reduced overall and conditional root mean standard error for all the study cases, with only slightly increased chi-square measure and test overlap. Compared to FSH, USTR greatly increased the use of low alpha items, decreased overexposure of high alpha items, and reduced test overlap while achieving comparable test efficiency. When the pool size was large enough relative to test length, FSH was found to result in higher efficiency, but its dramatically skewed item exposure distribution and high test overlap remained a problem. Preliminary results suggest that USTR might be a promising alternative. (Contains 3 tables, 6 figures, and 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 459 213

a-Stratified Computerized Adaptive Testing with Unequal Item Exposure across Strata

Hui Deng

The University of Iowa

Hua-Hua Chang

National Board of Medical Examiners

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

H. Deng

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033512

Paper presented at the 2001 Annual Meeting of the American Educational Research Association, Seattle, USA.

BEST COPY AVAILABLE

a-Stratified Computerized Adaptive Testing with Unequal Item Exposure across Strata

Computerized adaptive testing (CAT) has become an operational practice in many testing organizations due to its promise of increased measurement efficiency and precision. Critical to the realization of such advantages is its item selection algorithm. A commonly used algorithm is the maximum Fisher information procedure (Hambleton, Swaminathan, & Rogers, 1985; Lord, 1980), which proposes to select items that have the maximum information at each of a succession of trait estimates. This approach leads to great enhancement in test efficiency such that a CAT can achieve great precision with as few items as possible.

When applied in practical settings, the maximum Fisher information method is usually coupled with some nonstatistical constraints for validity considerations. One important constraint is item exposure control, which overrides the item selection algorithm to prevent any item from being overexposed, as a safeguard to test security (Way, 1998). Among a variety of item exposure control methods that have been proposed, the Sympson-Hetter procedure (Sympson & Hetter, 1985) is most widely used. This method attempts to control the conditional probability that an item is administered when selected, such that the observed exposure rate of any item falls less than or equal to some prespecified maximum value. The use of the Sympson-Hetter method coupled with maximum Fisher information (FSH) helps limit the exposure rates of items that otherwise would be overexposed, but it has been found to result in skewed distribution of item exposure rates. Specifically, some low *a* (item discrimination parameter) items had zero exposure rates, while many high *a* items achieved maximum exposure rates at the early stages of CAT (Chang & Ying, 1999; Hau & Chang, 1998). This is because Sympson-Hetter method has

no direct control over the probability that an item is selected, therefore those items with small probabilities of being selected may still be underexposed.

Chang and Ying (1999) proposed a new item selection methodology, the *a*-stratified multistage CAT (STR). The procedure starts with stratifying the item pool into same sized subpools according to item *a* values. Corresponding to these strata, the test is partitioned into multiple stages. At each stage, the same number of items are selected from the corresponding stratum based on the match of the item *b* (item difficulty parameter) value and the current trait estimate. The active stratum changes as the test proceeds such that low *a* items are used in the early stages and high *a* items are used in the later stages of the test. The key rationale for this method is that the low *a* items should be used in the early stage of the test since they provide more global information when trait estimates may be substantially deviant from the true level, while in the later stages high *a* items can better contribute to estimation when trait estimates are getting closer and closer to the true level (Chang and Ying, 1996). An apparent advantage of STR is its natural and effective balance of item usage. By grouping items with similar *a* values together and selecting within one group at each stage, items with all *a* values would be selected with about equal frequency. This method, however, has no direct control over the exposure rates for individual items, thus some items may still be overexposed unless an additional control algorithm is imposed. The benefits of combining the Sympson-Hetter algorithm and the STR design (STR-SH) has been studied (Leung, Chang, & Hau, 1999). The results suggest that the use of Sympson-Hetter effectively minimized the number of overexposed items, enhanced balancing usage of items and reduced test overlap rate.

The implementation of the STR method requires that items in each stratum cover a rich

spectrum of *b*-parameter values. When *a* and *b* values are correlated , however, this method may not function well. To remedy this problem, a simple moderation was proposed by Chang, et. al (2000) which adopts a two-stage stratification. Specifically, the item pool was first partitioned into blocks of equal size according to *b* values. Then within each block items were stratified into as many levels as number of strata needed according to *a* values. Finally items of the same level across blocks were combined to form strata. This method of *a*-stratification with pre-*b*-blocking was found to work successfully with pools which consisted of items with correlated *a* and *b* values.

STR has been compared to FSH with respect to item usage and test efficiency using both simulated and operational item pools (Chang & Ying, 1999; Hau & Chang, 1998). The results suggest that STR resulted in more evenly distributed item exposure rates and reduced overlap rate, while achieving slightly higher overall mean squared error (MSE) as compared to FSH. A close inspection of the Chang and Ying (1999) study reveals that, for 3000 simulated 20-item CATs with an operational pool of 254 items, the observed usage rates for most items were slightly below the level of 0.1. With the maximum exposure rate set to .20 for items in the pool, the above result suggests room for improvement of test efficiency. By drawing equal number of items from each strata, the current STR method may fail to make sufficient use of high *a* items in the later stages of CAT, thus degrading test efficiency to some extent.

It has been commonly realized that a tradeoff exists between test efficiency and item exposure control. A choice needs to be made that maximizes efficiency within the limits of security constraints, and that is essentially a matter of optimization. As a widely adopted view, a preferred item selection method should have evenly balanced usage rates with each item used

with about equal frequency (Mills & Stocking, 1996). This serves primarily for the convenience of item pool management. When test efficiency is considered, however, this may not be the optimal choice. To ensure high efficiency of a CAT, sometimes high *a* items should be used more than low *a* items. This strategy of varying item exposure according to item *a* values is expected to lead to improved test efficiency. Based on this consideration, a refinement of the STR method is proposed in this study that allows unequal item exposure across strata. By selecting more items from high *a* strata and fewer items from low *a* strata, it is anticipated that test efficiency will be enhanced with item usage rates that still follow a distribution not too far from a uniform pattern.

The purpose of this study was to explore the efficiency gain of the revised *a*-stratified method (USTR) and compare it with its original version STR and the FSH method with respect to test efficiency and item pool usage using simulated CATs. The comparison would be first made using a simulated item pool and then validated using a real item pool.

Method

Two simulation studies were conducted to compare the proposed USTR with its original version STR and FSH with respect to efficiency of trait estimation and item usage. In study 1, a set of item parameters were generated to form a bank of 400 items and the possible gains of using an USTR design were explored. In study 2, the comparison was carried out under a more complex design using a real item pool. The item parameters came from a large-scale standardized testing program, where items were calibrated using large samples of real data on mathematical test items using the three-parameter (3PLM) logistic item response model . In both studies, the 3PLM model forms the basis for all the data simulations.

Study 1: Simulated Item Pool

Item pool structure

The item pool consists of 400 items. The item *a* values were set at .5, 1.0, 1.5, and 2.0 for each of the four levels of 100 items. The *b* values were randomly generated from the standard normal distribution N(0,1). The *c* values were all fixed to zero. For the two stratification methods, the four levels of 100 items correspond to the four strata, with ascending *a* values across strata.

Design for CAT administration

With 400 items in the pool, the test length was set at 32 items to make the ratio of the pool size to test length 12, a number recommended as a rule-of-thumb by Stocking (1994). 3000 θ values were generated from the standard normal distribution N(0,1). Maximum likelihood estimation (MLE) was used to estimate θ for all CATs. To get an initial MLE, three artificial items were used. For the first item, a_1 was set to 1, b_1 was randomly selected form N(0,1), and c_1 was set to .2. If this item was answered correctly, the second item was made more difficult with $b_2 = b_1+2$, otherwise $b_2 = b_1-2$. *a* and *c* remained unchanged. The same procedure was used for the third item. Beyond the three initial items, item selection followed the procedures detailed below.

Item selection

For the STR method, 8 items were selected from each of the four strata corresponding to test stage, with each item being selected minimizing the absolute difference between *b* and the current $\hat{\theta}$.

For the USTR method, 4, 6, 10, 12 items were drawn from each stratum, adopting the same criterion as STR of matching b and $\hat{\theta}$.

For the FSH method, item i was selected if the following two conditions were met: (1) the item Fisher information has the maximum value at the current $\hat{\theta}$, and (2) $u<P(A_i/S_i)$, where u is a randomly selected number form a uniform distribution U(0,1) and $P(A_i/S_i)$ is the exposure control parameter. If condition (2) was not met, the next best item was evaluated; the process continued until both conditions were met. The exposure control parameters P(A/S) for all items in the pool were obtained through a series of iterative simulations following the steps specified below (Sympson & Hetter, 1985):

Step 0: Set P(A/S)=1 as initial values for all items in the pool.

Step 1: Administer the CATs. When an item i is selected by the maximum information criterion, generate a random number u from a uniform distribution U(0,1), administer the item only if $u \leq P(A_i/S_i)$, otherwise the next best item was considered. After all CATs are finished, compute the probability of the item being selected $P(S_i)=N_{si}/N$, and the probability of the item being administered $P(A_i)=N_{Ai}/N$, where N_{Si} is the number of times item i is selected, N_{Ai} is the number of times it is administered, and N is the number of examinees taking the test. Get P(S) and P(A) for all the items in the pool (Note that P(S)=P(A) after the first iteration).

Step 2: Adjust P(A/S) according to a pre-specified maximum exposure rate r. If $P(S)\leq r$ then P(A/S)=1, else if $P(S)>r$ then $P(A/S)=r/P(S)$.

Step 3: Repeat Steps 1 and 2 until all P(A/S)s have stablized and the maximum P(A) value approximately equals r. The P(A/S) values from the final iteration, one for each item in the pool,

are the exposure control parameters to be used in CAT administrations.

Evaluation criteria

To evaluate test efficiency, the overall bias and root mean square error (RMSE) of θ estimates were computed using the following equations:

$$Bias = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)^2}$$

where θ_i (i=1,2,...,m) denotes the true ability, and $\hat{\theta}_i$ the estimated ability for examinee i. m represents the number of examinees in the sample.

For evaluation of item usage, the following χ^2 (Chang &Ying, 1999) was used to measure discrepancy between the observed and the uniform exposure rates:

$$\chi^2 = \sum_{j=1}^{N}(er_j - \overline{er_j})^2 / \overline{er_j},$$

where er_j is the observed exposure rate, and $\overline{er_j}=L/N$ is the uniform exposure rate with L being the test length and N being the pool size.

Test overlap rate is another important summary index for tracking item exposure control (Way, 1998). It represents the expected proportion of items shared by a randomly chosen pair of examines. Specifically, the calculation involves the following steps: (1) counting the number of common items for each of the $m(m\text{-}1)/2$ pairs of examinees, (2) summing all the $m(m\text{-}1)/2$ counts, and (3) dividing the total counts by $Lm(m\text{-}1)/2$. To ensure test security, the number of common items shared by any two randomly sampled examinees should be minimized.

Study 2: Real Item Pool

The real item pool consists of 300 mathematical items with parameters calibrated from the real data. Unlike the simulated pool where the structure of item parameters can be arbitrarily determined to suit for a certain type of design, for a real item pool it should be the other way around: the characteristics of the items in the pool should be considered when planning for CAT design and item selection. For the items in this pool, the correlation between the *a* and *b* values was .356 (significant at .0001 level), calling for *b*-blocking being performed for the two stratification methods. Using the two-stage stratification procedure outlined in Chang et. al (2000), the pool of 300 items was partitioned into four strata with 75 items within each. The summary statistics for the item parameters are presented in Table 2 for the entire pool and for each stratum. As shown in the table, after stratification with pre- *b*-blocking, the means and SDs of *b* values appeared to be similar across strata, while the means of *a* values increase across strata.

To investigate the effects of test length and maximum exposure rate on the performance of the various methods, three cases of test conditions were examined.

Case 1 represents a situation where the available item exposure is rich relative to the requied exposure. In this case, the test length was set at 20 items and the maximum exposure rate was set at .20, thus the available item exposure, (300)(.20)m=60m (m is the number of simulees), was three times the required exposure, 20m. For the STR method, 5 items were selected from each stratum; while for the USTR method, 3, 4, 6, 7 items were selected from each of the four strata, respectively.

Case 2 represents a condition of more demanding required exposure, with test length

increased to 32 items and the maximum exposure rate remained .20. Here the available item exposure, still $60m$, was a little short of twice the required exposure, $32m$. For the STR method, 8 items were selected from each of the four strata; while for the USTR method, 4, 6, 10, 12 items were selected from each stratum, respectively.

In Case 3, the test length was equal to that in Case 1, whereas the maximum exposure rate was reduced to .10. This represents a very stringent security requirement. In this case the available item exposure, $(300)(.10)m=30m$, was just one and a half times the required exposure, $20m$. The specifications of numbers of items to be selected form each stratum remained the same as in Case 1.

To mimic the real world situation where constraints on item exposure need to be satisfied, the Sympson-Heater precedure (SH) was incorporated into the STR and/or USTR methods if many items were found overexposed, following the procedure detailed in Leung et. al(2000).

To provide a comparison for the estimation efficiency conditional on ability level, RMSEs were examined for the various methods at each of the eleven points equally spaced on the θ scale, ranging from -2.5 to 2.5 in increments of 0.5. At each point 1000 simulees were generated for CAT simulations.

The other aspects of the CAT design, the item selection and the evaluation criteria remained same as in study 1.

Results

Study 1

Table 1 displays the performance of the three item selection methods in terms of ability estimation and item pool usage. The bias measures were small for all methods, indicating the three methods were all virtually unbiased. FSH had the smallest RMSE .141, while STR had the largest, .170. Note the RMSE for USTR was .152, much smaller than that for STR, suggesting that allowing more items to be selected form high *a* strata was effective in improving test efficiency.

Concerning item usage, FSH had the highest χ^2 value, 36.344, over five times the value for STR (6.898) and over three times that for USTR (11.433). Since the χ^2 measure was used here to capture the discrepancy between the oversued and the ideal uniformly distributed item exposure rates, the above results suggest that FSH resulted in the most unevenly distributed rates. Figure 1, showing exposure rates against item number ordered by *a*-parameter, provides more details. For the FSH method, many items with high *a* values approached or exceeded the maximum rate, while items with low *a* values showed near zero exposure rates. In fact, as Table 1 shows, among the 400 items in the pool, 178 items had exposure rates below .02 when FSH was used, and there were 4 and 34 such items for STR and USTR, respectively. Figure 1 also illustrates the different patterns of exposure rates for the two stratification methods. The graph for STR showed a more balanced pattern, but most items were used with rates between .05 and .10. When USTR was used, however, the usage rates for the low *a* items were decreased and the rates for the high *a* values were increased, as expected. Compared to STR, the less evenly distributed exposure rates

of USTR contributed to its increased χ^2 measure, as a sort of trade-off for the increased efficiency.

As to the overlap rate, FSH had the highest value, .171, which was 5.5 (.171*32=5.5) overlapping items on average for a 32-items test for any two randomly sampled examinees. While for STR and USTR, the average number of overlapping items were 3.1 and 3.5 respectively. Varying item exposure across strata led to a slight increase of test overlap rate.

Overall, the results from study 1 suggest that the use of USTR effectively improved test efficiency, with a price of slightly increased unbalance of usage rate and slightly increased overlap rate. Given the substantively increased test efficiency , it seemed warranted to further study the method using a real item pool.

Study 2

The relative performance of the three selection methods were studied under three cases which involve different combinations of test length and maximum exposure rate. The results for each case are summarized below.

Case 1

Table 3 presents the summary of performance for the three methods for all cases. In case 1, with the test length fixed to 20 and the maximum exposure rate set to .20, the pool of 300 items provided much selection space for the CATs. As seen in Table 3, FSH appeared to be superior in terms of test efficiency as it had the lowest overall RMSE, .291. USTR had the lowest bias (.002), although FSH and STR also had small biases. FSH yielded largest χ^2 value (33.379), almost six times that of STR (5.490) and five times that of USTR (6.936), indicating

very unbalanced item usage. FSH also had the highest overlap rate (.178) among the three methods. Compared to STR, USTR showed improvement on efficiency. Its RMSE measure (.324) was smaller than that for STR (.345), yet it resulted in slight increases on χ^2 and overlap rate. The above results were in much agreement with those for study 1, as a similar design was employed in the two conditions. Note the ratio of pool size to test length was 15 in this case, characterizing an even richer item availability.

Figure 2 displays the item exposure rates plotted against items ordered by *a* values for this case. FSH showed a very unbalanced distribution, with many low *a* items underused and high *a* items approaching maximum exposure rate. Many points can be seen located on the horizontal axis, indicating zero exposure rates for those items. In addition, some items with middle or high *a* values exceeded the maximum rate, demonstrating a shortcoming of Sympson-Hetter procedure in that some items may still be overexposed even with the algorithm in function. The pattern of exposure rates for STR appeared most even among the three methods, but several items exceeded the maximum rate .20. While the graph for USTR showed somewhat more variability of exposure rates compared to STR, no item exceeded the maximum rate. For both STR and USTR, no points were seen on the horizontal axis, indicating all items were used to some extent.

To further examine the characteristics of item exposure for the three methods, the frequency distributions of exposure rates were plotted and shown in Figure 5. Based on the criterion for optimal item exposure (Wainer, 2000), the ideal distribution should be peaked and unimoded, indicating most or all the items are exposed at roughly the same frequency. As seen in

the top graph which demonstrates the results for case 1, STR appeared to be best with respect to balanced item usage, with nearly 37% of the items used at a rate between .04 to .06 and none of the items was used below a rate of .02. With USTR, about 27% of the item in the pool were exposed at rates between .02 and .04, and another 20% of the items were exposed at a rate between .06 and .08. The situation for FSH was much worse, with 45% of the items never exposed, and 12.5% of items with rates exceeding .20.

Figure 6 showed RMSEs conditional on ability levels for the three methods. As seen in the top graph relevant to case 1, FSH resulted in much lower RMSEs than the other two methods at all ability levels except the lower extreme of the scale, and the discrepancies were particularly salient in the middle range. USTR also yielded lower RMSEs than STR along the ability scale except at the high end, although the difference was not as pronounced as for FSH.

To summarize, the results for case 1 indicated that, when the size of the item pool and the maximum exposure rate provided a rich availability of item exposure, the use of FSH can achieve considerable efficiency gain relative to stratification methods, yet with the price of very unbalanced item exposure. The USTR resulted in improved efficiency over STR with small compromises on balanced item usage and overlap rate.

Case 2

In case 2, the maximum exposure rate was still .20 but the test was lengthened to 32 items, representing a more demanding situation in terms of required item exposure. Since some items were found overexposed using USTR (as can be seen in Figure 3), an additional method (USTR-SH) with Sympson-Hetter incorporated into USTR was examined. As shown in Table 3, all the four methods appeared unbiased, and STR had the highest overall RMSE. FSH still had

the lowest RMSE, but USTR and USTR-SH had very similar RMSE values as FSH. The χ^2 value for the FSH , 23.388, was highest among the four methods, but smaller than in case 1 (33.379). The higher demand of required item exposure seemed to force more balanced exposure rates. Again, the use of USTR resulted in increased χ^2 value and overlap rate, as trade-offs for improved efficiency. The use of Sympson-Hetter with USTR effectively brought the overexposed items under the maximum (see Figure 3). It also further improved efficiency (as shown by the RMSE value), reduced the χ^2 value, and achieved the same overlap rate as compared to USTR.

From the middle graph in Figure 5, it can be seen that the distribution of exposure rates for STR was most peaked, with 30% of the items used at a rate between .06 and .08. ; while the curves for UATR and USTR-SH appeared flatter. For FSH, 23% of the items in the pool were never used, a smaller proportion compared to case 1, suggesting lengthening the tests brought more items to use. However, more items were overexposed than in case 1 (24% compared to 12.5%), also due to the effect of lengthening the tests.

The middle graph in Figure 6 presents the results for the conditional RMSEs for case 2. FSH had the lowest RMSE only at the middle portion of the ability scale, and the difference was much smaller than in case 1. In the ability range of -2 to 0, USTR had lower RMSEs than FSH. USTR-SH showed similar RMSEs as USTR along ability scale, except in the range below zero where the curve for USTR-SH appeared slightly higher. In general, both USTR and USTR-SH showed lower RMSEs than STR along the ability scale, particularly in the middle range of the scale. The efficiency gain with the use of USTR (and USTR-SH) seemed more pronounced than

in case 1, so did its departure from the uniform distribution of exposure rates. This was because the pattern of item numbers across strata (4, 6, 10, 12) allowed more high *a* items to be fully used than in case 1, which can also be observed by comparing the USTR graph in Figure 2 with the USTR-SH graph in Figure 3.

In general, the results from case 2 suggest that the efficiency advantage for FSH was reduced when CATs were lengthened (Note in this case, the ratio of pool size to test length was about 10). The use of USTR, especially with Sympson-Hetter incorporated seemed a viable option. Compared to FSH, it ensured test efficiency while at the same time improved balanced item usage balance and reduced test overlap.

Case 3

In case 3, the available item exposure was constrained by setting a stringent maximum rate, .10. The test length remained 20 items as in case 1. Since many items were found overexposed for both STR and USTR methods, the Sympson-Hetter procedure was incorporated into both methods, resulting in two more methods (STR-SH and USTR-SH) for comparison. As shown in Table 3, USTR yielded the lowest overall RMSE (.318), and STR and STR-SH the highest (both .351). The use of SH with USTR increased RMSE to .340, approaching the value for FSH (.339). FSH still had the highest χ^2 value and overlap rate, but not as much higher than other methods as in cases 1 and 2. Imposing a low maximum for item exposure (.10) seemed to have restricted the use of many items and forced the exposure rates to become more balanced. The use of Sympson-Hetter reduced χ^2 values and overlap rates for both STR and USTR cases, and the χ^2 values and overlap rates for the two USTR methods were slightly higher than their

STR counterparts.

As shown in the bottom graph in Figure 5, the distribution of item exposure rates for STR was peaked at rates between .04 to .05. For FSH, about 21% of the items in the pool were still never utilized, and about 25% of items were used with rates exceeding the maximum rate .10. Setting a stringent maximum rate helped to bring more items to use compared to case 1, but still one-fourth of the items in the pool were never touched.

From the bottom graph in Figure 6, it can be seen hat FSH yielded the lowest RMSEs only at the ability range between 0 to 1.5, otherwise the curve for USTR was the lowest along the scale. USTR-SH had similar RMSE values as USTR in the ability range above zero, and higher values in the range below zero. The curves for STR and STR-SH were consistently close to each other along the ability scale.

The results from this case suggest that when a stringent maximum exposure rate was imposed, FSH again lost its advantage in efficiency. In addition, item exposure rates for all selection methods were forced to become more uniform, and overlap rates were reduced. While FSH still resulted in many items either never used or overused, USTR (with Sympson-Hetter as needed) seemed to offer a solution for overcoming the item usage problem and at the same time maintain test efficiency.

Discussion

In this study a revised *a*-stratified item selection method, USTR, was proposed and compared with the unrevised version STR and the traditional method FSH using CATs simulated from both artificial and real item pools. The results suggest that the USTR holds considerable

potential for improving test efficiency for STR. It effectively reduced overall and conditional RMSE measures for all the study cases, with only slightly increased χ^2 measure and test overlap. Compared to FSH, USTR greatly increased usage of low *a* items, decreased overexposure of high *a* items, and reduced test overlap while achieving comparable test efficiency. When the pool size was large enough relative to test length, FSH was found to result in higher efficiency, but its dramatically skewed item exposure distribution and high test overlap remained a problem.

In a recent article, Wainer (2000) demonstrated a nonlinear relationship between item usage and item security. It was clearly shown in the article that to maintain the security level of a test over time, there must be exponentially more items added to the pool, which is a very discouraging finding as far as test development and item pool maintenance are concerned. In fact, using the traditional method FSH, what is running out of the pool quickly are those items with high *a* values. To maintain an item pool in order to support effective CATs, items with high *a* values need to be continuously input to the pool. At the same time those items must satisfy test specification and other constraints. This is a very difficult task for item writters, which even if achievable, may not be economically feasible at all. A solution is needed.

The preliminary results from the current study suggest that USTR might be a promising alternative. It, to a large extent, overcomes the item usage problem for FSH while at the same time restores test efficiency. Even though with this method high *a* items may run out a little quicker than if STR is used, it maybe worthwhile in terms of the increased efficiency. One might argue that the efficiency gain of using USTR over STR may not be desirable under all circumstances considering the compromised item usage. We agree with this concern and recommend that the

choice of method should be based on a informative judgement with many factors taken into account, such as test purpose, test stakes, security requirements, and very importantly, the characteristics of the items in the pool.

The simulations conducted in this study were based on a specific set of real items and limited in scope, thus the results obtained are rudimentary and far from complete. More studies are needed to extend the work using other pools in order to offer a more definitive picture of the strengths and weaknesses of the USTR method. It should also be noted that, content constraints were not considered in this preliminary study, therefore future research should address the effects of content constraints on the comparative performance of USTR and other methods. Another direction for future work is how to specify the pattern of different item numbers across strata. Depending on the structure of the item pool, different patterns can be tried out and compared using simulations during CAT design stage. The optimal pattern for a USTR design is expected to be related to the distribution of a-parameters in the pool. In addition, the number of strata and the size of the pool should also be taken into consideration. Future research is needed to determine the relationship among these factors and provide general guidelines for implementation of USTR in real world settings.

References

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20, 213-229.

Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23, 211-222.

Chang, H., Qian, J., & Ying, Z. (2000). A-stratified multistage CAT with *b*-blocking. Applied Psychological Measurement, in press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hau, K., & Chang, H. (1998). Item selection in computerized adaptive testing: Should more discriminating items be used first? Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Leung, C., Chang, H., & Hau, K. (1999). Item selection in computerized adaptive testing: improving the a-stratified design with the Sympson-Hetter algorithm. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, CA.

Lord F. M. (1980). Applications of item response theory to particle testing problems. Hillsale, NJ: Lawrence Erlbaum.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. Applied Measurement in Education, 9, 287-304.

Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools (ETS Research Report No. 93-2). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized

adaptive testing. Proceedings of the 27th Annual Meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's law. Journal of Educational and Behavioral Statistics, 25, 203-224.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. Educational Measurement: Issues and Practice, 17, 17-27.

Table 1. Performance Summary for the Various Methods for Study 1

Method	N	*r*	Bias	RMSE	χ^2	#<.02	Overlap rate
FSH	32	0.2	0.004	0.141	36.344	178	0.171
STR	32	0.2	0.003	0.170	6.898	4	0.097
USTR	32	0.2	0.000	0.152	11.433	34	0.108

Note. N: Test Length

r: Desired maximum exposure rate

Pool size: 400 items

Table 2. Summary Statistics for the Item Parameters for the Real Item Pool

Stratum	# of Items	Item Parameter	Mean	S.D.	Minimum	Maximum
1	75	*a*	0.696	0.184	0.296	1.114
		b	0.312	0.972	-2.348	2.582
		c	0.160	0.046	0.042	0.267
2	75	*a*	0.911	0.202	0.487	1.407
		b	0.303	0.967	-2.515	2.199
		c	0.151	0.040	0.073	0.252
3	75	*a*	1.072	0.205	0.493	1.505
		b	0.308	0.939	-2.132	2.005
		c	0.145	0.053	0.031	0.271
4	75	*a*	1.259	0.247	0.584	1.933
		b	0.310	0.937	-2.185	1.971
		c	0.137	0.043	0.037	0.251
Total	300	*a*	0.984	0.295	0.296	0.193
		b	0.308	0.949	-2.515	2.582
		c	0.148	0.046	0.031	0.271

Table 3. Performance Summary for the Various Methods for Study 2

	Bias	RMSE	χ^2	Overlap rate
N=20, r =.20				
FSH	0.006	0.291	33.379	0.178
STR	-0.009	0.345	5.490	0.085
USTR	0.002	0.324	6.936	0.089
N=32, r =.20				
FSH	0.001	0.258	23.388	0.184
STR	-0.002	0.279	5.888	0.126
USTR	0.001	0.262	11.152	0.144
USTR-SH	-0.004	0.260	10.025	0.144
N=20, r =.10				
FSH	-0.007	0.339	8.960	0.096
STR	-0.002	0.351	5.500	0.085
STR-SH	-0.010	0.351	3.080	0.077
USTR	0.010	0.318	6.751	0.089
USTR-SH	-0.003	0.340	4.491	0.081

Note. N: Test Length.

r: Desired maximum exposure rate.

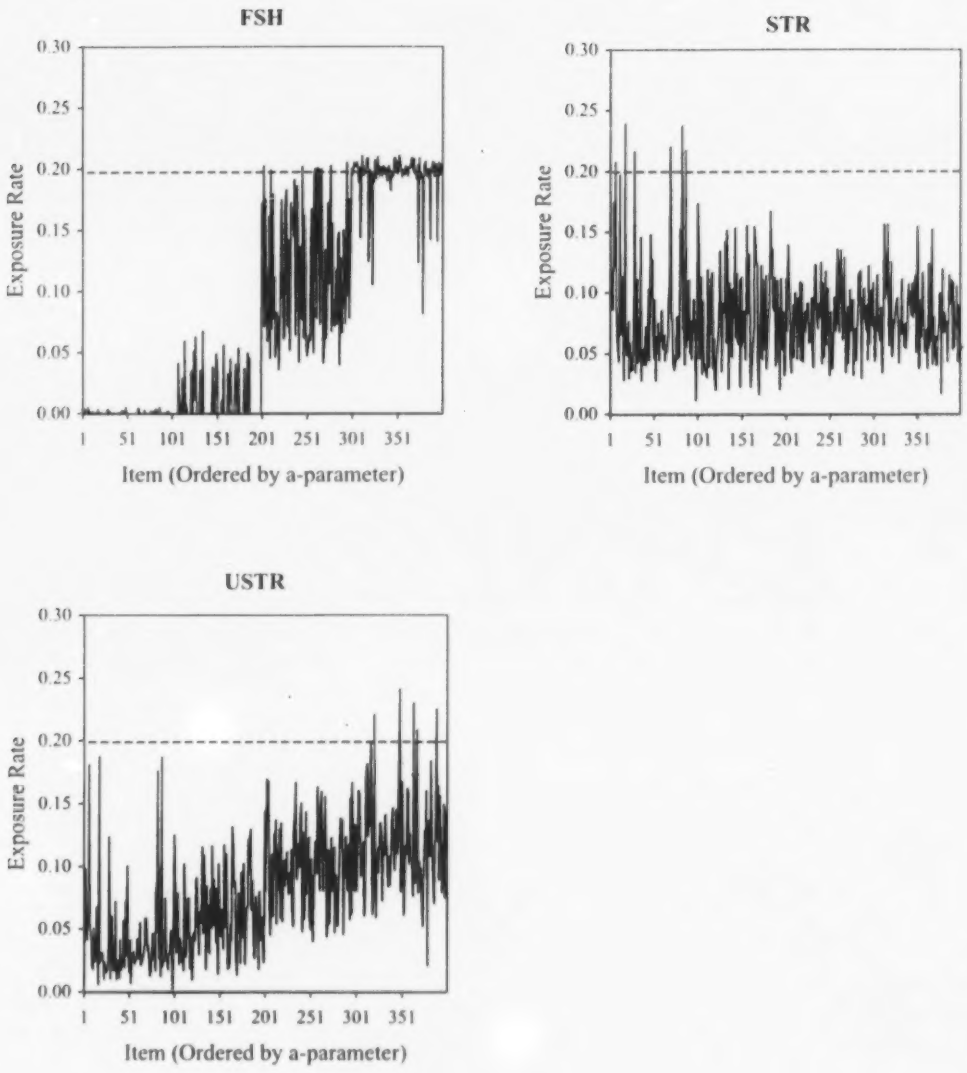

Figure 1. Item Exposure Rates for 3000 Simulated 32-Item CATs (with r =.20) for Study 1

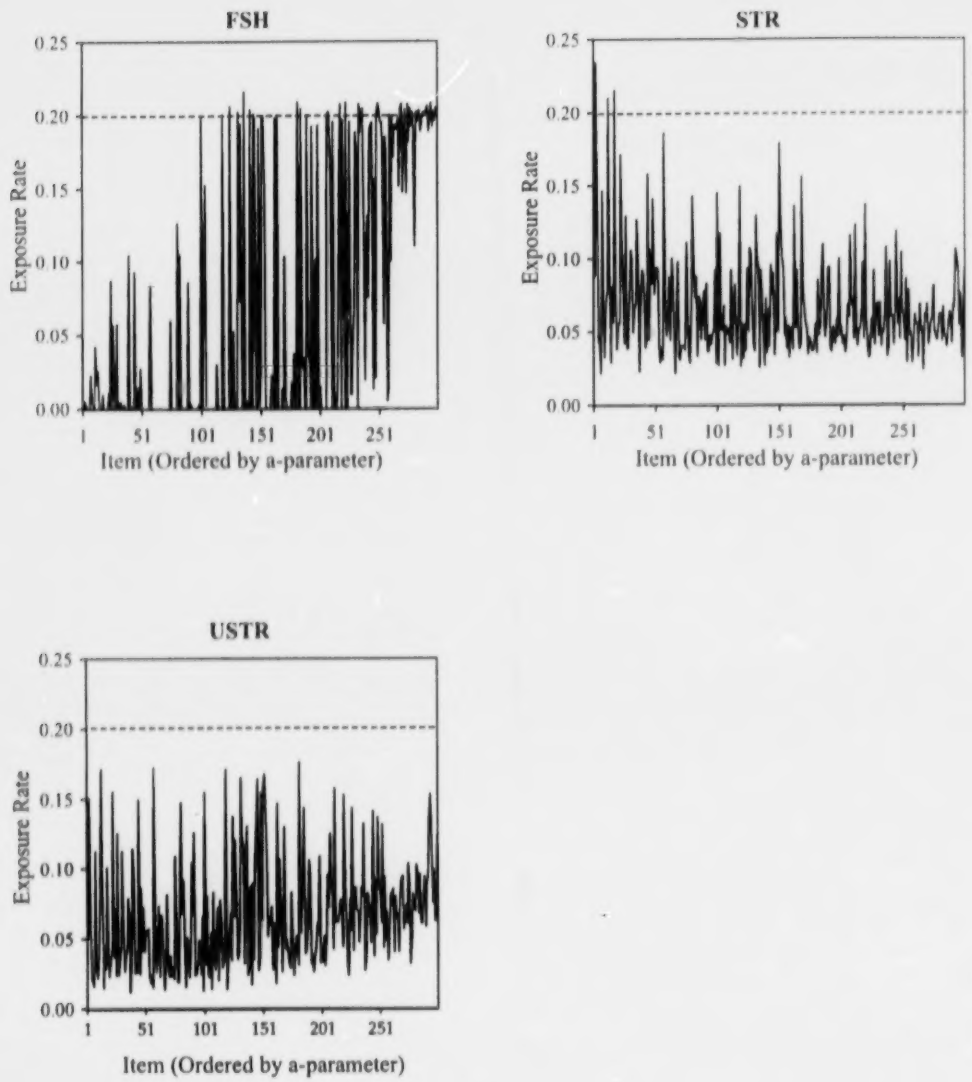

Figure 2. Item Exposure Rates for 3000 Simulated 20-Item CATs (with r = .20) for Study 2

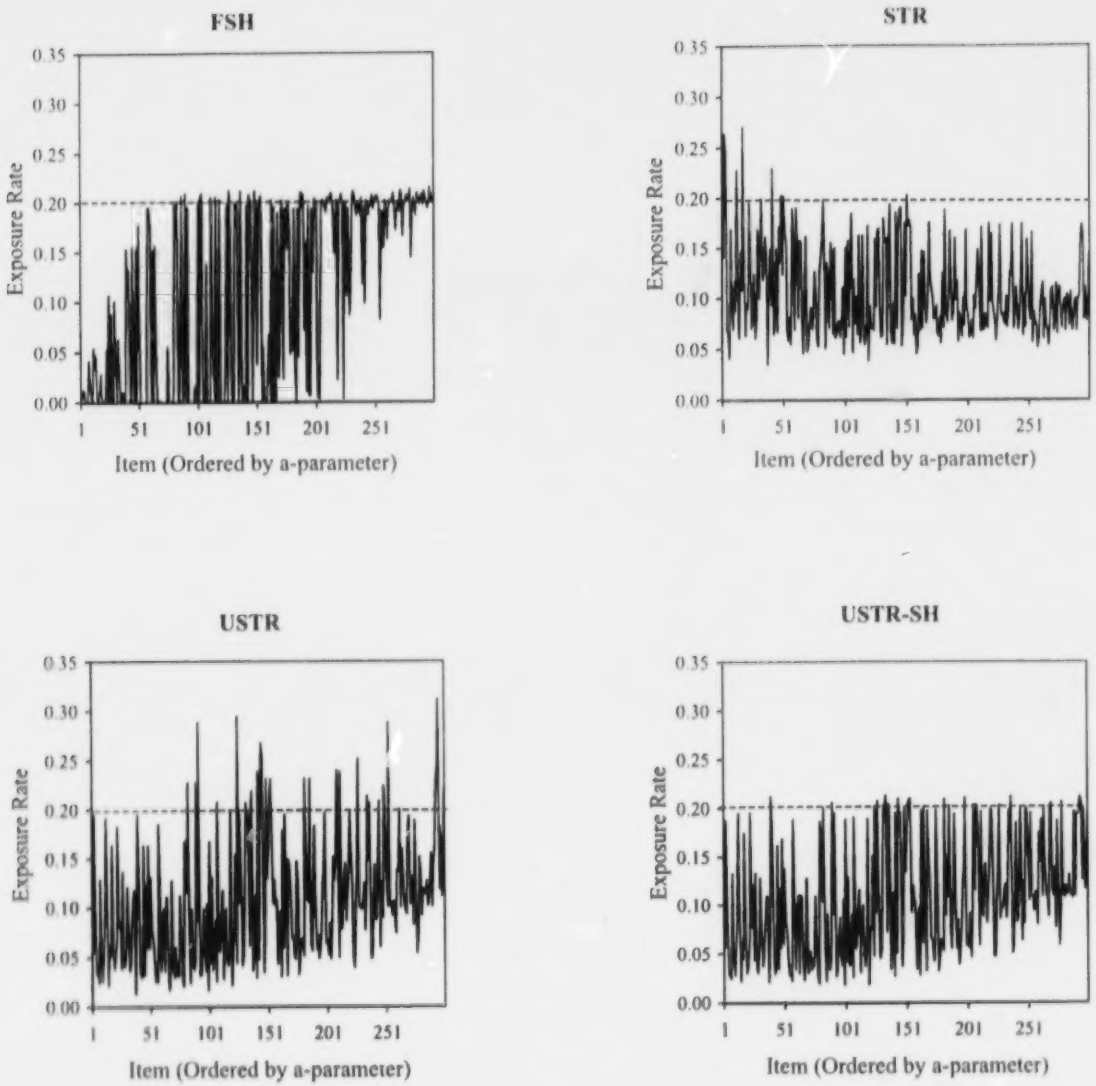

Figure 3. Item Exposure Rates for 3000 Simulated 32-Item CATs (with r = .20) for Study 2

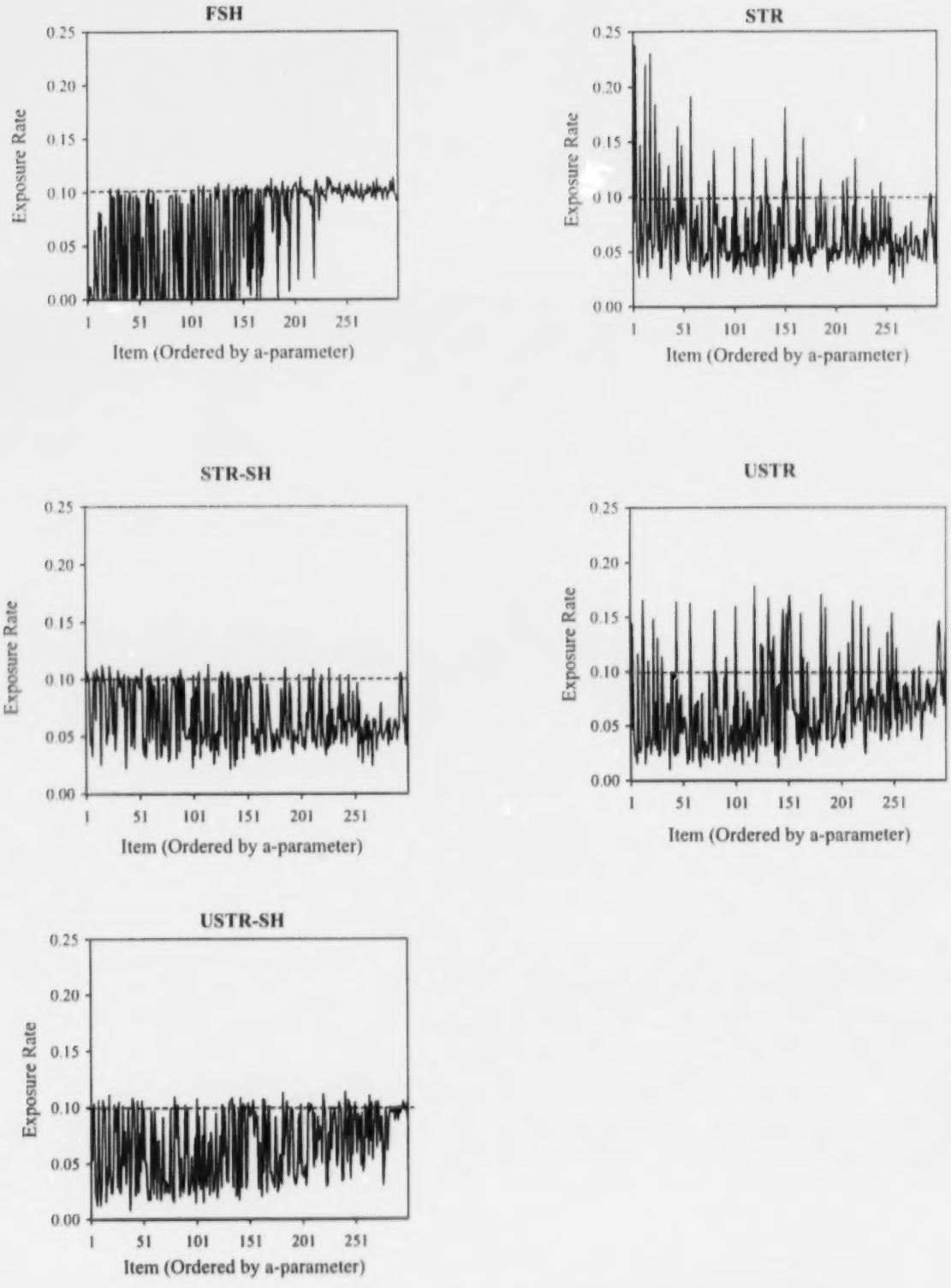

Figure 4. Item Exposure Rates for 3000 Simulated 20-Item CATs (with r =.10) for Study 2

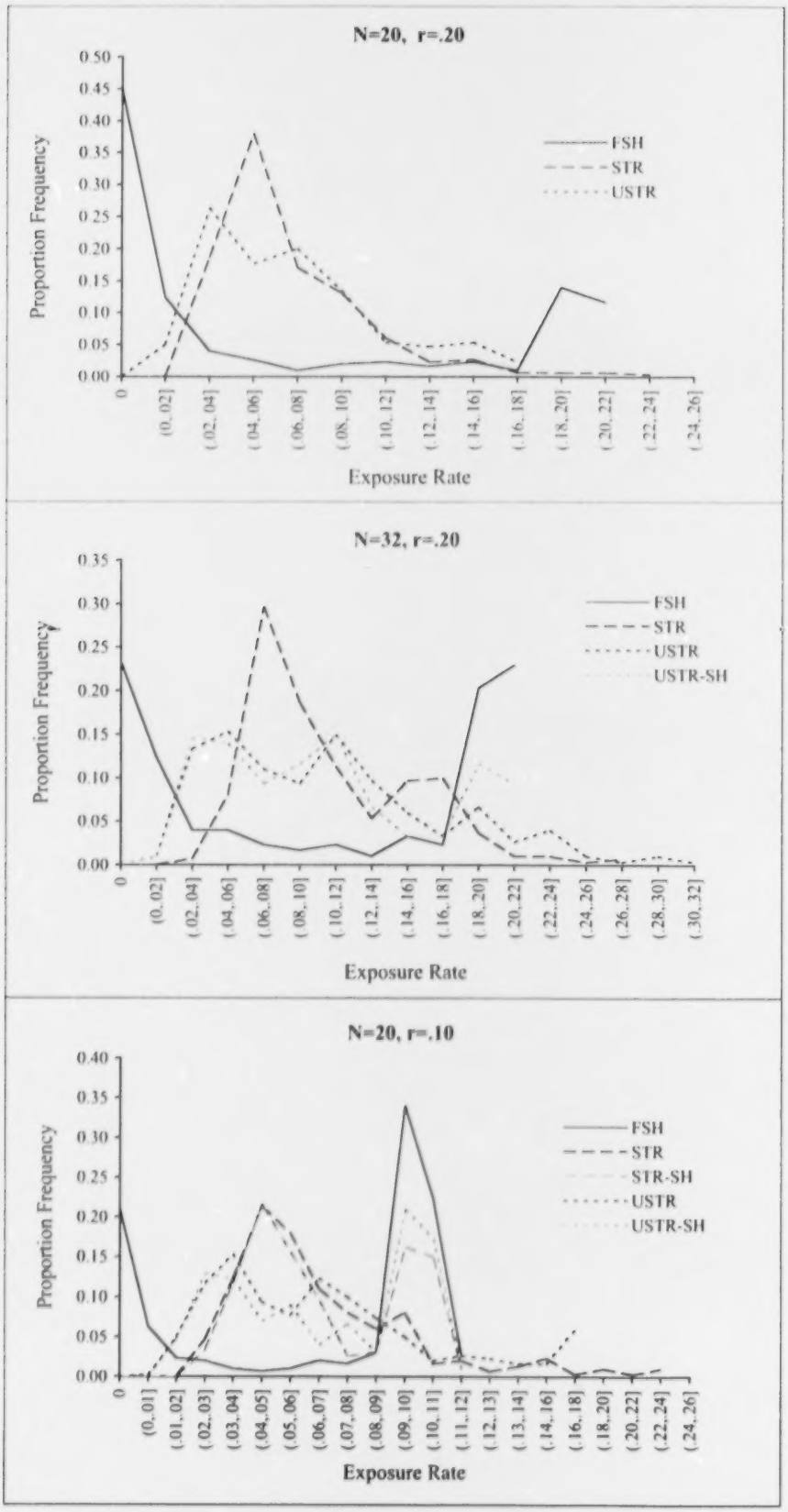

Figure 5. Distribution of Item Exposure Rates for the Various Methods for Study 2

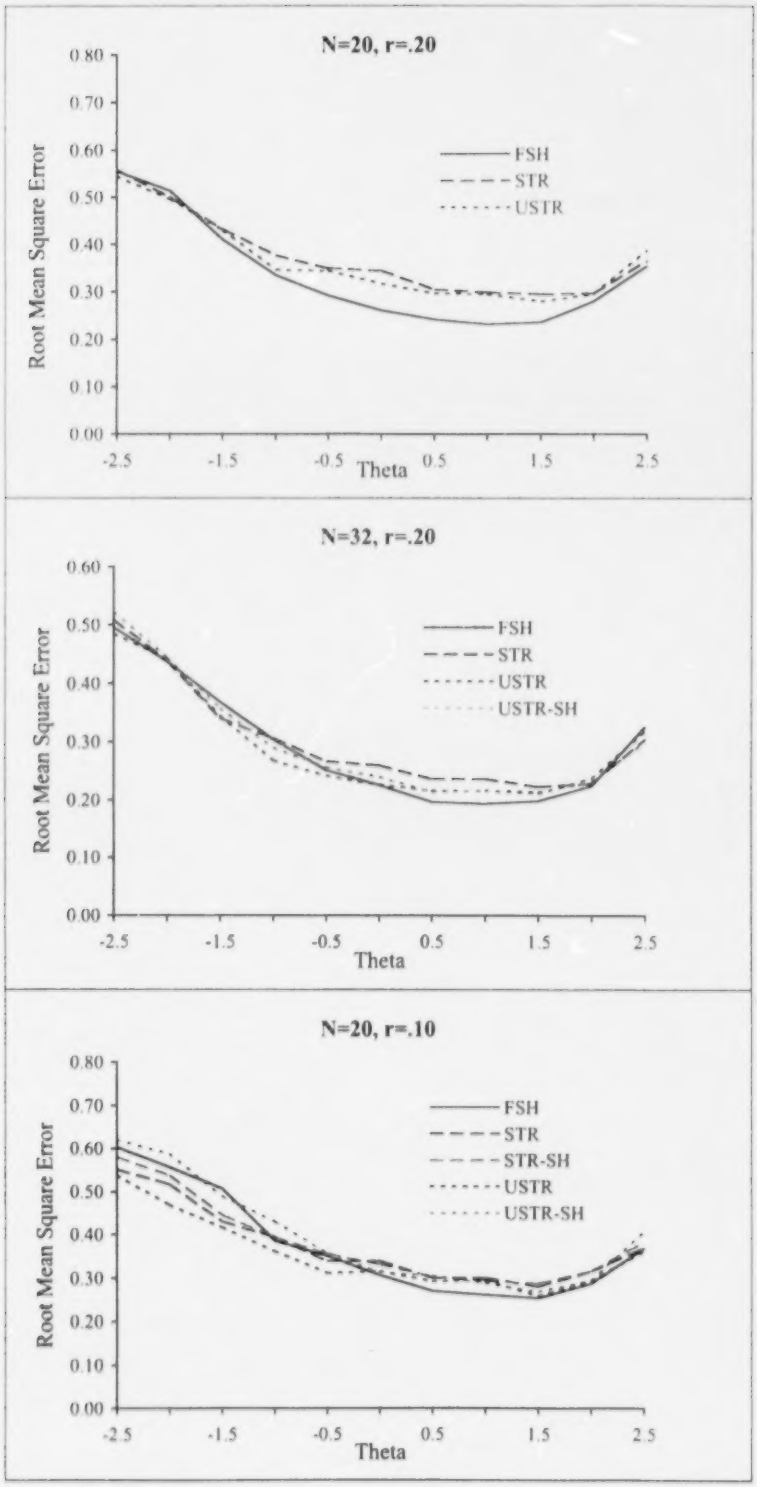

Figure 6. Conditional Root Mean Square Error for the Various Methods for Study 2